haberman_dataset

October 23, 2018

Haberman Data Set

The Haberman dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone surgery for breast cancer.

Attribute Information:

```
<ol>
```

Age of patient at time of operation (numerical)
Patient's year of operation (year - 1900, numerical)
Number of positive axillary nodes detected (numerical)
Survival status (class attribute)

```
  1 = the patient survived 5 years or longer
 </li><li>
  2 = the patient died within 5 year
 </li></ol>
</li>
</ol>
```

Axillary lymph nodes and breast cancer

Sometimes, breast cancer can spread to the axillary lymph nodes, which are in a person's armpits.

The number of axillary lymph nodes can vary from person to person, ranging from 5 nodes to more than 30.

When someone is diagnosed with breast cancer, knowing if cancer has spread to their axillary lymph nodes can determine the type of treatment they have, as well as their prognosis (an estimate of the future, especially about whether a patient will recover from an illness. [formal] If the cancer is caught early the prognosis is excellent.).

```
In [1]: import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        import numpy as np
        %matplotlib inline

        df = pd.read_csv('haberman.csv', names = ['Age', 'Op_Year', 'axil_nodes_det', 'Survive
        df.head()
```

```
Out[1]:    Age  Op_Year  axil_nodes_det  Survived_5_years_or_longer
        0   30       64               1                           1
        1   30       62               3                           1
        2   30       65               0                           1
        3   31       59               2                           1
        4   31       65               4                           1
```

Survived_5_years_or_longer: convert the datatype to bool
'2' = False and '1' = True

```
In [2]: df['Survived_5_years_or_longer'] = df['Survived_5_years_or_longer'].apply(lambda x : T
        df.head()
```

```
Out[2]:    Age  Op_Year  axil_nodes_det  Survived_5_years_or_longer
        0   30       64               1                        True
        1   30       62               3                        True
        2   30       65               0                        True
        3   31       59               2                        True
        4   31       65               4                        True
```

```
In [3]: df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 306 entries, 0 to 305
Data columns (total 4 columns):
Age                          306 non-null int64
Op_Year                      306 non-null int64
axil_nodes_det               306 non-null int64
Survived_5_years_or_longer   306 non-null bool
dtypes: bool(1), int64(3)
memory usage: 7.5 KB
```

```
In [4]: df.describe()
```

```
Out[4]:              Age     Op_Year  axil_nodes_det
        count  306.000000  306.000000      306.000000
        mean    52.457516   62.852941        4.026144
        std     10.803452    3.249405        7.189654
        min     30.000000   58.000000        0.000000
        25%     44.000000   60.000000        0.000000
        50%     52.000000   63.000000        1.000000
        75%     60.750000   65.750000        4.000000
        max     83.000000   69.000000       52.000000
```

Understanding the Data

75% of the people who were treated for cancer are below the age of 60.75 years, while the lowest and highest age were 30 and 83 respectively

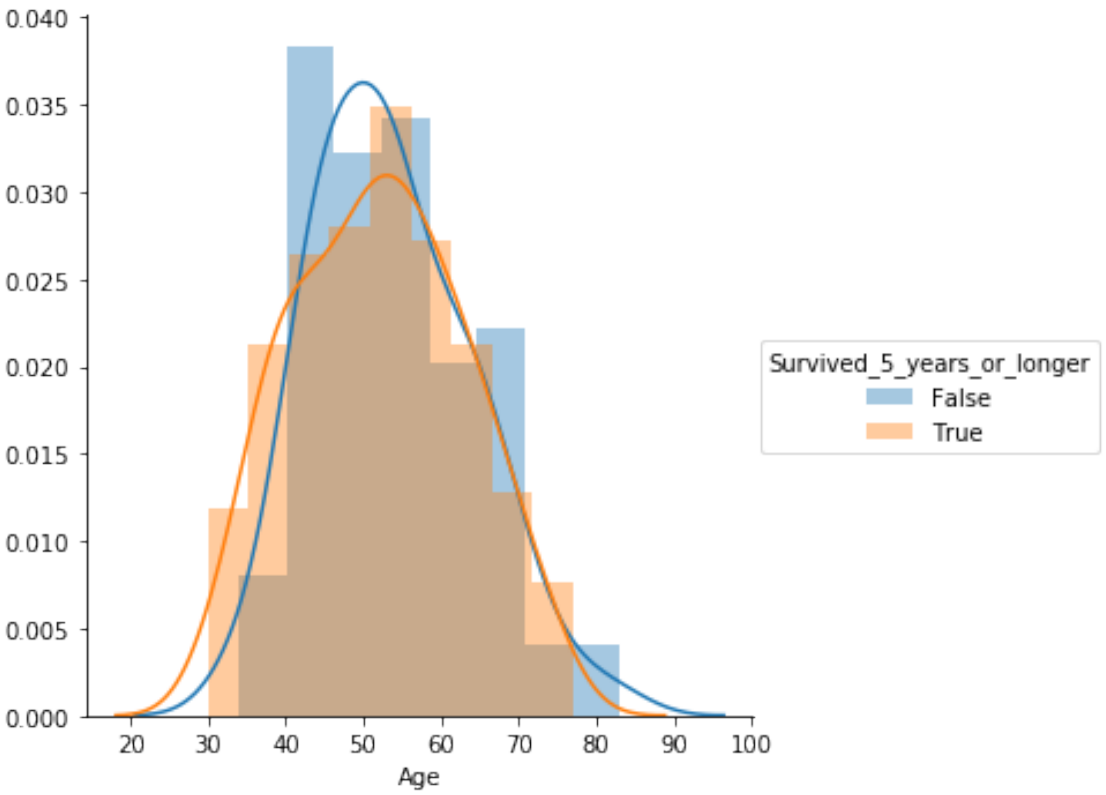
75% of the people who were treated had less than 5 positive axillary nodes detected with a high of 52

```
In [5]: grouped = df.groupby('Survived_5_years_or_longer').count()

In [6]: print("Number of Points ={}, Number of Features={}, Number of Classes={}".format(df.sha
        print("Points Per Class=")
        df["Survived_5_years_or_longer"].value_counts()

Number of Points =306, Number of Features=3, Number of Classes=2
Points Per Class=


Out[6]: True     225
        False     81
        Name: Survived_5_years_or_longer, dtype: int64
```

haberman dataset is an imbalanced dataset as the number of survivors for 5 years or longer is more then number of people who died following the surgery in less than 5 years

UniVariate Analysis

```
In [7]: #PDF, CDF, BoxPlots, Violin Plots
        sns.FacetGrid(df, hue="Survived_5_years_or_longer", size=5) \
           .map(sns.distplot, "Age") \
           .add_legend();
        plt.show();

/anaconda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tu
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
/anaconda/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6571: UserWarning: The 'normed'
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```

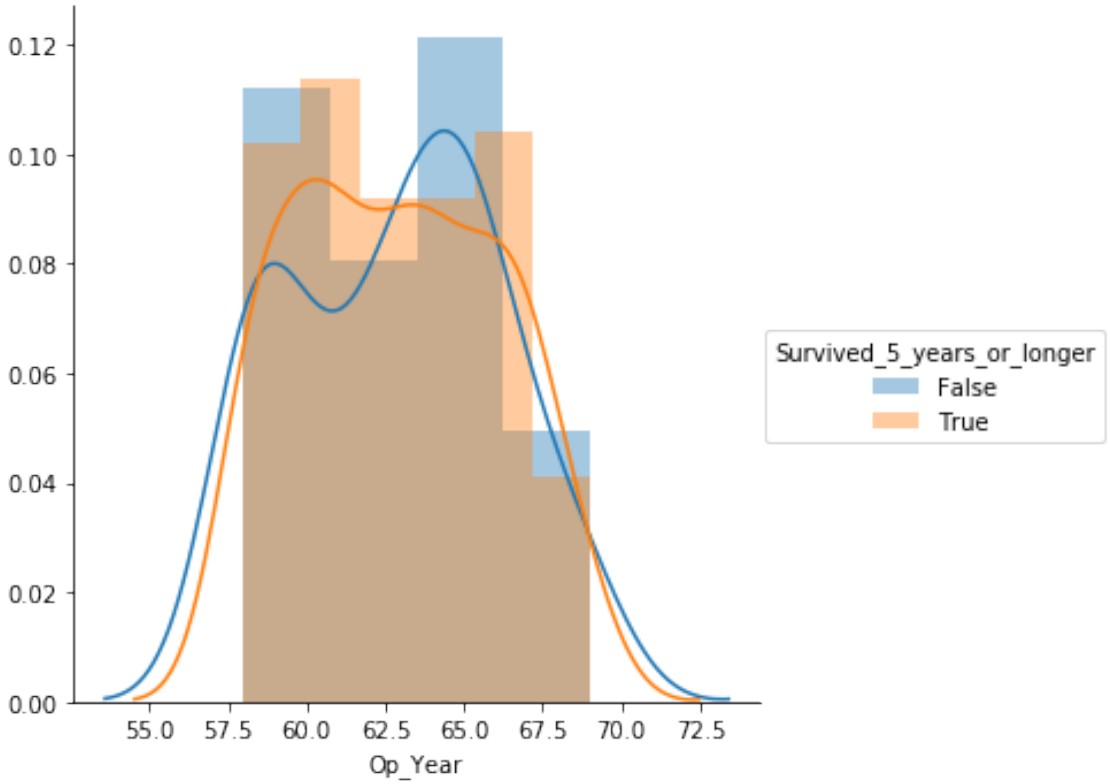


Observation There is a large overlapped region in the Histogram of Age and hence its not easy to define a simple model for classifying the survival based only on Age. For an Age in the large intersection area between 35 to 75 one cannot decisively say if Survived_5_years_or_longer is true or false

```
For the Age  between 30-35 the survival indicator is strong, potentially indicating that the cl
</p>
```

```
In [8]: #PDF, CDF, BoxPlots, Violin Plots
        sns.FacetGrid(df, hue="Survived_5_years_or_longer", size=5) \
           .map(sns.distplot, "Op_Year") \
           .add_legend();
        plt.show();
```

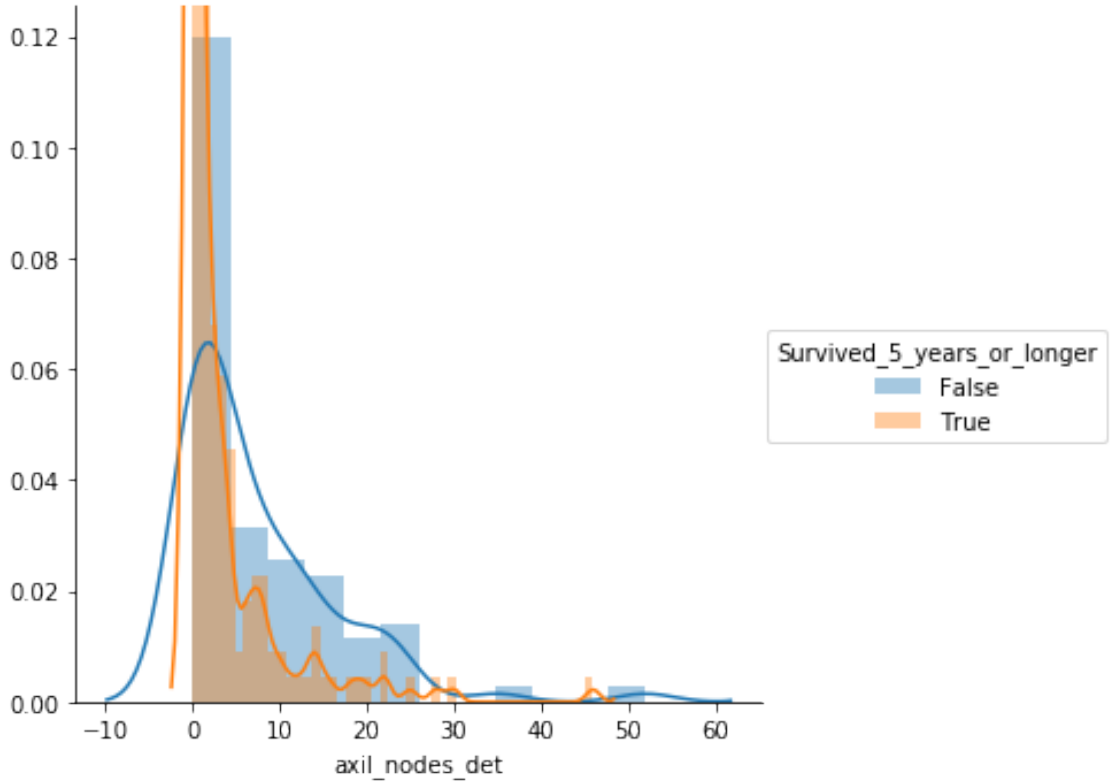
```
/anaconda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tu
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
/anaconda/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6571: UserWarning: The 'normed'
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



Observation There is a large overlapped region in the Histogram of Operation Year and hence its not easy to define a simple model for classifying the survival based only on Operation Year. For an Operation Year in the large intersection area between 57 to 68 one cannot decisively say if Survived_5_years_or_longer is true or false

```
In [9]: #PDF, CDF, BoxPlots, Violin Plots
        sns.FacetGrid(df, hue="Survived_5_years_or_longer", size=5) \
           .map(sns.distplot, "axil_nodes_det") \
           .add_legend();
        plt.show();
```

```
/anaconda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tu
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
/anaconda/lib/python3.6/site-packages/matplotlib/axes/_axes.py:6571: UserWarning: The 'normed'
  warnings.warn("The 'normed' kwarg is deprecated, and has been "
```



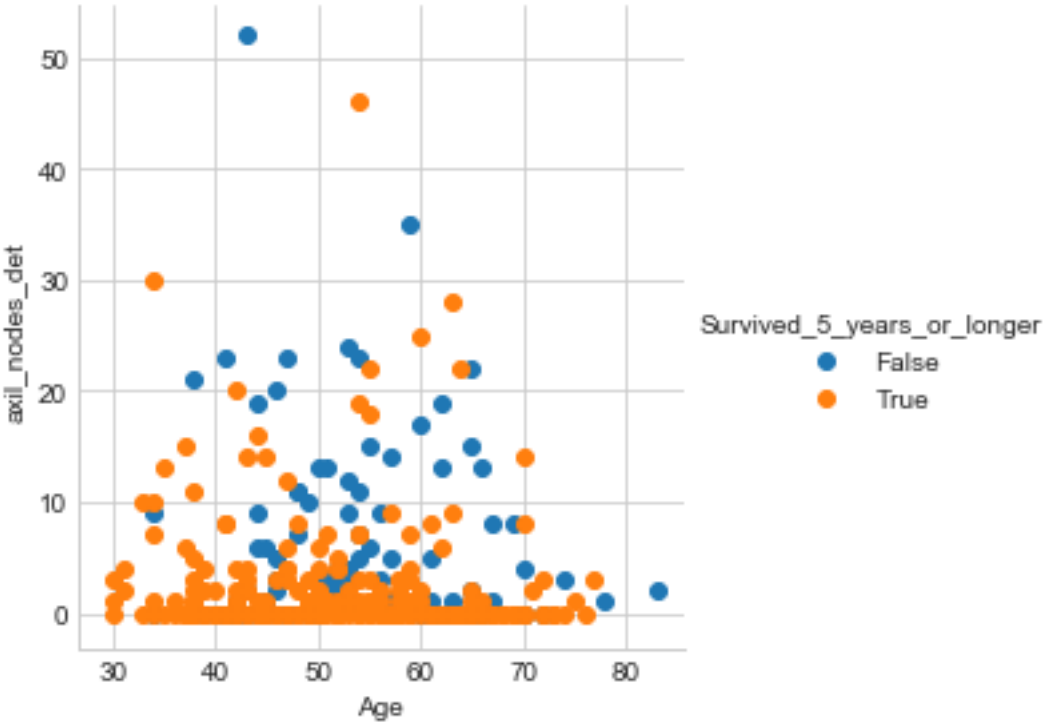
Observation Unlike Age and Operation Year the PDF for axil_nodes_det does not appear to be approximately normal and seems to have a right tail.

```
<li>
For axil_nodes_det between 5 and 25, the number of deaths is more then the number of survivals
</li>
<li>
  The number of survivals reduces rapidly as the value of axil_nodes_det increases beyond 5.
</li>
<li>
    There are outliers around axil_nodes_det value of 25-30 and around 45 where the data shows
</li>
</ul>
```

2D Scatter Plot

2D Scatter plot using Age and axil_nodes_det.

```
In [9]:
        sns.set_style("whitegrid");
        sns.FacetGrid(df, hue="Survived_5_years_or_longer", size=4) \
           .map(plt.scatter, "Age", "axil_nodes_det") \
           .add_legend();
        plt.show();
```



Observations

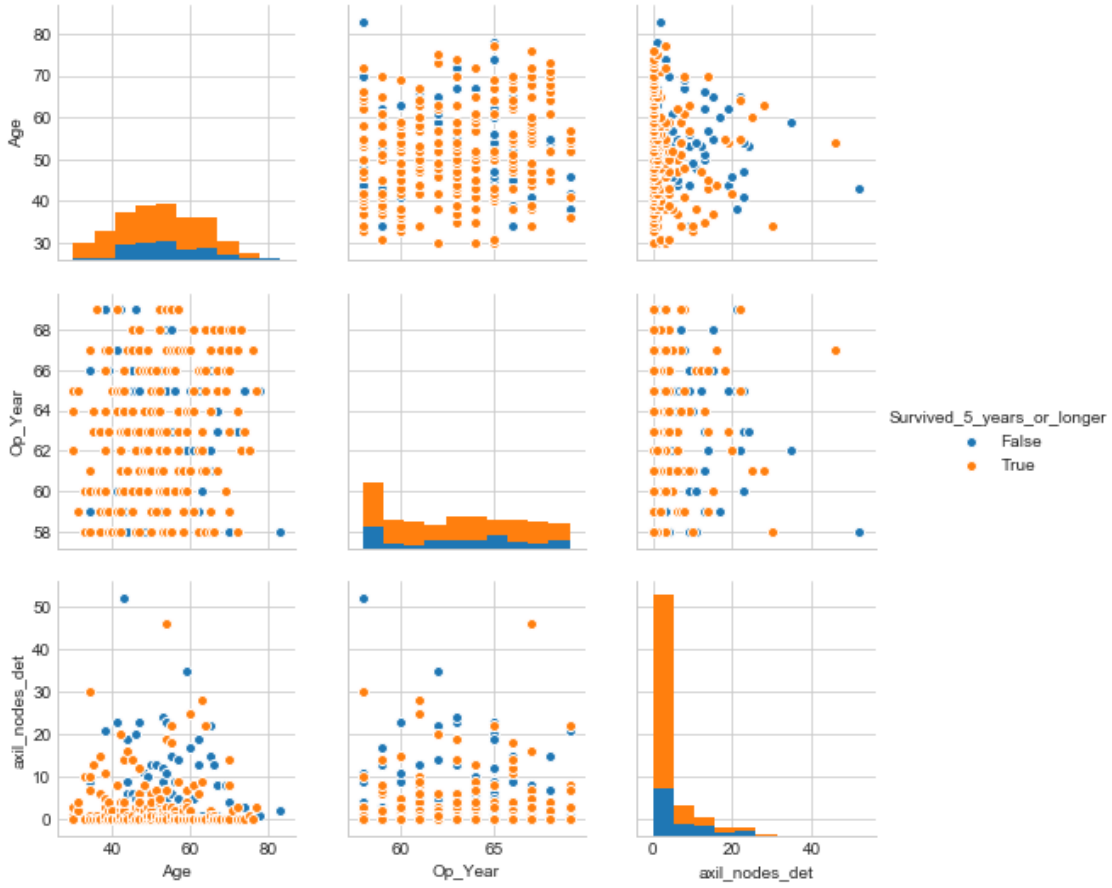
The Survival rate is highest when the number of positive axillary nodes detected is low (close to zero)

But the 2 classes cannot be easily seperated in this 2D scatter plot.

Need to try Pair Plots for more analysis

Pair Plots

```
In [12]: plt.close();
         sns.set_style("whitegrid");
         sns.pairplot(df, hue="Survived_5_years_or_longer", x_vars=["Age", "Op_Year", "axil_noc
         plt.show()
```



Observations

```
<li>
  The Pair Plot : X=Op_Year and Y=Age, does not seem to show any specific relationship between
</li>
<li>
  The Pair Plot : X=axil_nodes_det and Y=Op_Year, does not seem to show any specific relationsh
</li>
<li>
     The Pair Plot : X=Age and Y=axil_nodes_det, shows that for higher ages the survival rate i
</li>
<li>
  There is one data point to suggest that if the Age is above 80 then survival rate is low irre
</li>
```

```
In [13]: plt.close();
         sns.set_style("whitegrid");
         sns.pairplot(df, hue="Survived_5_years_or_longer", x_vars=["Age", "Op_Year", "axil_noc
         plt.show()
```

```
/anaconda/lib/python3.6/site-packages/scipy/stats/stats.py:1713: FutureWarning: Using a non-tu
  return np.add.reduce(sorted[indexer] * weights, axis=axis) / sumval
```